DOCUMENT RESUME

ED 166 250 TM 008 290

AUTHOR Kennedy, Dale
TITLE The Nebraska Blueprint for School Assessment Programs. Volume 1: Purposes and Categories of Measurement.
INSTITUTION Nebraska State Dept. of Education, Lincoln.
PUB DATE 78
NOTE 48p.; For related documents, see TM 008 291 and 292

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Achievement Tests; Aptitude Tests; *Educational Assessment; *Educational Testing; Elementary Secondary Education; *Guidelines; Intelligence Tests; Interest Tests; Performance Tests; *Testing Programs; Tests

ABSTRACT

This is the first of three volumes designed to cover the problems and issues of developing and managing a formal assessment program. They also offer guidelines and illustrations of several aspects of assessment programs, such as administering tests or documenting their results. This volume deals with the regulations governing school assessment programs in Nebraska, the general purposes for which assessment information is collected, and the kinds of measuring instruments used for such collection. It provides a summary of existing measurement theory and measurement technology with respect to educational settings. It outlines general requirements for assessment and special requirements for Title I and for programs for handicapped children. It discusses the purposes of assessment for accountability, diagnosis, and management and criterion measures. It discusses varieties of tests and assessment techniques under three headings: general ability (intelligence or general aptitude and general achievement or educational development); special abilities (primary mental abilities and aptitude batteries and various achievement measures); and measures of typical performance, specifically interest inventories. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED166250

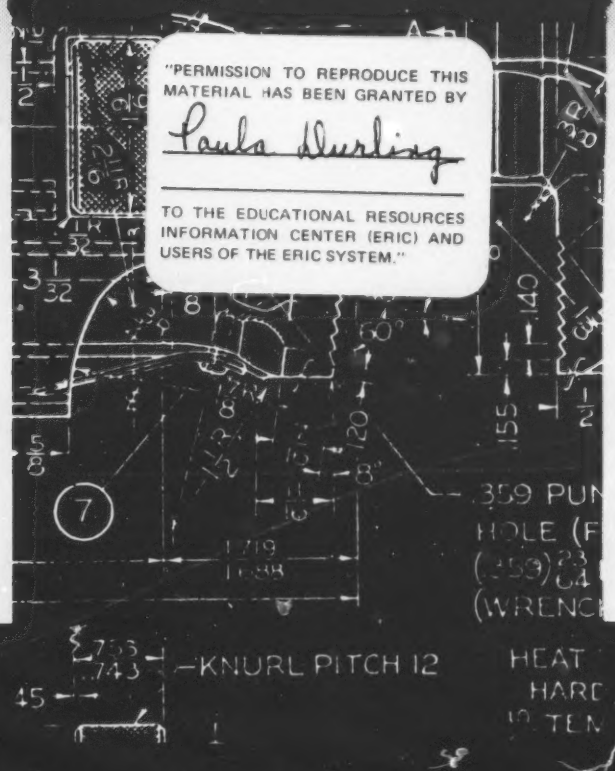
U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Paula Hurling

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."



The Nebraska

BLUE PRINT

For School Assessment Programs

Volume 1: Purposes and Categories of Measurement

Issued by the Nebraska State Department of Education

M. Anne Campbell, Commissioner of Education

301 Centennial Mall South
P. O. Box 94987
Lincoln, Nebraska 68509

1978

TABLE OF CONTENTS

Volume I: Purposes and Categories of Measurement

1. Introduction 1
   - General Requirements 1
   - Special Requirements 3
     - Title I 4
     - Special Education 5
   - Responsibility 6

2. General Purposes 7
   - Progress and Passages: The Criterion Measures 8
   - Accountability 9
   - Diagnosis 11

3. The Varieties of Tests and Assessment Techniques 13
   - I. General Ability 14
     - A. Intelligence or General Aptitude 14
       - Evaluation Criteria 15
     - B. General Achievement or Educational Development 18
       - Evaluation Criteria 19
   - II. Special Abilities 20
     - A. Primary Mental Abilities and Aptitude Batteries 20
       - Evaluation Criteria 22
       - Reading Readiness 23
     - B. Special Achievement 23
       1. Conventional Standardized Tests 23
          - Evaluation Criteria 24
       2. Criterion-Referenced Measurements 27
          - Evaluation Criteria 29
       3. Teacher-Made Tests 30
       4. Applied Performance Testing 30
          - Performance Sampling 31
          - Evaluation Criteria 32
          - Simulation 33
          - Applied Performance Batteries 34
   - III. Typical Performance 34
     - Interest Inventories 35

COMMISSIONER'S FOREWORD

Assessment, in ordinary words, is the effort devoted to finding out if students are learning what we think we are teaching. Like instruction, it is a complex undertaking, both in conception and in practice. Both aspects, instruction and assessment, are essential to the process of education and both require careful and thorough development to be of best effect. The Department's Blueprint for School Assessment Programs covers the problems and issues of developing and managing an assessment program. It also offers guidelines and illustrations of several of the more special aspects of assessment programs, such as administering tests or documenting their results. The Blueprint's coverage is comprehensive and will have value for a variety of audiences that may be involved in assessment programming.

It is anticipated that the Blueprint will prove a useful reference in planning, developing, evaluating, reviewing, and managing local assessment programs.

M. Anne Campbell
Commissioner of Education

AUTHOR'S FOREWORD

The Blueprint is intended to be a repository of theoretical and practical information on student assessment and the problems of assessment programming. It has the qualities each of a textbook, a handbook, and a reference manual. It is possible for it to be read and used at several different levels. It could satisfy readers having general interest in assessment theory and technology, but it should also prove to be of repeated value for readers who are actually involved in assessment programming, such as the planning committee members, examiners, and the persons charged with the school or district's assessment responsibilities. It may be scanned for general information or it may be studied for precise and technical information. It can serve to generally refresh and improve one's overall understanding of the issues and problems of assessment, or it can be employed as a guidelines and procedures manual for the actual planning and operation of an assessment program.

Those persons who are actually involved in the planning and operation of a school or district assessment should seriously consider acquiring and reading Standards for Educational and Psychological Tests, published by the American Psychological Association, Guide For School Testing Programs, published by the National Council on Measurement in Education, and Essentials of Psychological Testing, by Lee J. Cronbach, published by Harper and Row. There are other good books on psychological testing besides Cronbach's. The Essentials, however, exhibits the unique combination of well explained theory along with thorough discussions of practical concerns. Cronbach also writes in a scientific style

that is both exacting and conservative in statement. These are qualities which the Blueprint also attempts to achieve, though in a more synoptic form. The Blueprint, together with these three references, should soundly provide a thorough and comprehensive knowledge of assessment theory and practice.

Dale Kennedy

PREFACE

Human children are growing, learning, maturing, thriving organisms. The ingredients of their perceptions and cognitions, of their learning, far exceed those of other living things found upon earth, for children's growth must take place on several levels simultaneously: physical, intellectual and emotional. The purpose of education, then, is to shape and coax this growth (if not to stimulate it outright). Education, when harnessed to growth, leads it to its most mature levels in a way that ideally enhances further self education. An education "completed" ideally leaves the student open for a multitude of future circumstances.

Learning takes place as a function of three things: 1) past learning, in which one can include native ability because its effects are in fact manifested by the past learning; 2) the occasion upon which the learning takes place, i.e., the learning opportunity to the teacher and the learning experience to the student; and 3) the consequences contingent upon the learning behavior itself. Learning occurs if the behavior of the students takes a new shape during and after the occasion.

The change is lasting or strong, occurring again on appropriate future occasions or affecting future learning, in proportion to the reinforcement of the learning on the occasion of first change. But the student must first perceive the change and the consequences thereof. If the consequences are natural, as in the discovery of the world of literature through reading, the student must value these new possibilities. Teachers must also perceive the change if their actions toward the students are to be consistent, timely, and in pace with the students'

rate of growth. This reading of student behavior and ability is assessment.

Most assessment is informal, taking place each day, any time the teacher and student interact. Most of the student's self-assessment also takes place on informal occasions. The assessment function becomes formal when the student is asked to perform on occasions which would not normally be encountered as a daily routine, designed to reveal something of the student's ability. These occasions of formal assessment are the subject matter of the Blueprint.

The Blueprint is published in three separate volumes. Volume I, Purposes and Categories of Measurement, deals with the regulations governing school assessment programs, the general purposes for which assessment information is collected, and the kinds of measuring instruments used for such collection. This is the most conceptually involved volume, but it does provide a comprehensive view of existing measurement theory and measurement technology with respect to educational settings.

Volume II, Processes and Procedures, deals with planning of assessment programs, the program's standing documentation, and problems of administering, scoring, record-keeping and reporting, referred to collectively in the Blueprint as the logistics of management. The information provided in this volume balances the conceptual content of Volume I with detailed exposition on practical considerations. Volumes I and II are intended to complement each other.

Volume III, Technical Information, is nothing more than a repository of useful information of a technical or abbreviated nature. It has sections concerning reliability and validity, addresses of test publishers,

lists of references, computing guides for various statistics, test tables and a glossary. The information in this volume is also of a practical nature, even more so than Volume II.

There are four central assumptions underlying all volumes of the Blueprint.

First, a district's assessment program already exists and stands in a working relationship with the operations of instructional delivery. The primary object of the Blueprint is to provide the concepts and procedures that will bring the documentation of the program to a comprehensive, organized form, if it does not already exist in such a systematic fashion.

Second, the assessment of individuals and groups must be fair to those individuals and groups. If a test discriminates against an individual on the basis of background or physical handicap, the results will not represent true information on the individual's abilities or achievement. If used to make decisions about the individual's future education, such unfairly obtained data would be misinformation causing mistakes in judgment that may result in the miseducation of the individual. The fairness stricture also applies to mainstream groups. Here, if the administrations are poorly controlled, the margin of error is enlarged and the test becomes an inaccurate measure mostly containing misinformation. Assessments must be exactly administered if they are to be fair. Miseducation is the consequence of mistaken judgment.

The possibilities of miseducation are indeed some of the central reasons for establishing and maintaining an assessment program. For example, it is always desirable to fit instruction to known strengths

and weaknesses of the student, or groups of students. The educative process would be expected to reinforce the strengths, improve the weaknesses, and in its combination of effects stimulate an optimum amount of learning. But just as surely as instruction can promote optimum growth, it can drop a ceiling on learning. Instruction can mismatch, as well as match, the strengths and weaknesses of students. And in so doing, it fails to promote growth where potential is great or expects more of the student than the student is truly capable of. A mismatch guarantees a growth far less than what is optimum and, in effect, blocks such growth. Miseducation is simply the diminishment of learning through an inappropriately applied (indeed, an alleged) educative process. It is always a danger. Good assessment information should reduce and limit occurrences of this plight.

Third, grade promotion and graduation sanctions, contingent upon assessment information, are not necessary to effective and useful assessment programs. They are essentially a special credit system extended beyond the usual credits for attendance and passing courses. Where they are instituted, they became a defining feature of a whole assessment program, yet a program has several good uses without them. The logic and rationale of grade promotion and graduation requirements are not topics of the Blueprint. The subjects that are discussed, however, would give a soundness to the assessment information, so that if used for promotion, the information would at least be validly obtained.

Fourth, the readers of this manual and all those involved in assessment, either as tester, subject, or interpreter, should understand and appreciate the scientific nature of educational assessment. An interpre-

tation made from assessment information is a work of scientific reasoning. Both the proponents and the critics of scientific method as applied to education tend to regard the mechanical qualities of the methods of science as its only virtue, or weakness, as the case may be. This neglects the deliberateness and natural conservatism of scientific method. It also misses the cultivation of balanced perspective. Whether very mechanically obtained, or more informally so, assessment information must: first, be the right information needed; second, be accurate; and third, be viewed in relation to other assessment information on the same individual or group. If the information does not fit the questions asked, or if it is error-ridden, then it is misinformation and, if interpreted, would foster mistaken judgments and lead to the probable miseducation of the students. It requires an understanding of scientific method to recognize when and under what conditions the information gathered for the purposes of assessment proves to be misinformation, data which can be harmful to the future educational prospects of students.

Above all else, this scientific perspective must be grounded in a concrete understanding of the conditions under which the data are actually produced. It is the children who face the test and perform as best they can that day. The temperature of the room, the lighting, their breakfast or lack of breakfast, and many other daily, mundane facts, all conspire to affect the children's performance. The mechanical rigors of administration are intended to even out the disadvantages and to heighten the individual's performance. A testing procedure cannot be too precise when these are its objects.

General Requirements

Teaching has two inseparable aspects: the delivery of instruction and the assessment of learning. With respect to instruction, teaching involves the organization and management of the physical environment in which the learning experiences are to take place, the formation of instructional strategy (including materials and equipment to be used), and the actual conduct of the learning opportunities planned and arranged to stimulate student learning of a specific sort. With respect to assessment, teaching involves the diagnosis of the learning needs of individual students and groups of students, and the evaluation and reporting of student progress.[1]

Each of the different levels of operation, the teacher, the department, the school, and the district, has responsibilities to the teaching function. The teacher, of course, has the closest relation to the student, and delivers the instruction, by intent or default. The teacher also is in the best position to make the informal observations of the student's behavior that flesh out formal assessment. The department, though, also has clear teaching responsibilities, since collectively it will lay down the broader objectives of the curriculum under its jurisdiction and will have special interest in the group per-

---

[1]Rule 14, Revised, Regulations and Procedures for Approving the Contimed Legal Operations of All Schools and the Opening of New Schools, Approved by the State Board of Education, Effective August 11, 1975, Revised May 7, 1976, A2-3.

formances of these students passing through its domain. The school and district levels are extensions of jurisdiction to larger units or groupings of students. Each of these levels is subject to policy and procedures which govern the assessment that takes place at that level.

Section 8-(d) of Rule 11, "Rules and Regulations for the Accreditation of Public and Non-Public Schools," states:

> "Each school system shall have written policies which shall include but not be limited to the purposes for which group or individual assessment is done, procedures for the administration of the tests and procedures for its interpretation and dissemination of testing information."

Where assessment is conducted, its purpose, its method, interpretation and distribution as reported results must be decided upon and documented. The whole of such assessments is considered the assessment program of the district.

In particular, Rule 14, "Regulations and Procedures for Approving the Continued Legal Operations of All Schools and the Opening of New Schools," states that:

i. "The school system shall utilize a competency measure which assesses mastery of a minimum acceptable performance in reading, writing, and arithmetic skills." This measurement is to take place at both elementary and secondary levels. [14-(12) and 14-(50)].

ii. The cumulative record kept for each student must contain information on the student's academic progress [14-(48)]. There should be a written policy for the evaluation of academic progress [14-(49)]. This information is to be communicated to the student's parents. This also requires a written policy describing the process through which the communication is made [14-(51)].

The cumulative file of the individual student is the central unit of record. It should contain summaries of all assessment information collected on the student: group test scores; course grades; special

tests; and other documentation of accomplishment or special performance. At least some of the information collected will result from applied performance tasks in the skills of reading, writing and computation. Students must meet or exceed minimum standards of performance before the task is considered passed or mastered.

The written policies concerning purposes and procedures for administration, interpretation and dissemination of the assessment information constitute the documentation of the assessment program. This documentation, which includes the purposes of the assessment, the choice of tests, their scheduling, their specific uses and intended audiences, represents the standing record of the planning committee's (or advisory council) critical decisions regarding the assessment program. In brief, it is a composite summary of nearly all the decisions made to develop the program.

With full documentation, a district will be able to:

1) State and justify its purposes in assessment;

2) State and justify choices of specific tests and assessment techniques;

3) State and justify its assessment design;

4) State its procedures for the administration of its tests and assessment techniques;

5) State its procedures for scoring and recording the test and its results; and

6) Explain the manner in which the results are reported, to whom and to what end.

Special Requirements

If the district or school qualifies as a target area for funds for

special instructional needs, such as under Title I or Special Education, there are more particular regulations affecting the assessment program.

Title I - If a school qualifies under Title I criteria, it must comply with the following sequence of assessment.

1. In the first spring after qualifying, the school will administer a school-wide, standardized achievement battery. Students scoring below the 50th percentile are identified. These students are now eligible for special instructional programming. The achievement battery is to have norm group data for both spring and fall administrations.
2. In the fall that follows, a second, but different standardized achievement battery is administered, either school-wide or limited to the students identified in the spring testing. Previously identified students who remain below the 50th percentile become the target students. This second testing is confirmation of the first tests' results.
3. In the spring, after the delivery of the special programming, another form of the same achievement battery administered the previous fall is administered again. The battery is then changed again and the cycle begins anew.
4. The gains or losses of the target students between the fall and the spring test are reported on the Student Data Form of the State Title I Office and submitted to the same.

This sequence follows the Model A assessment guidelines of the U.S. Office of Education, developed especially for Title I evaluation. Models

B and C are more complex, but admissible. Information about the requirements of Models B and C is available at the Title I Office at the Nebraska State Department of Education.

Special Education - The particular assessment requirements of a special education program are presented in Rule 51, "The Development of Reimbursable Local Special Education Programs For Handicapped Children," section (4), passages 5.15 to 5.22. Special Education assessment is an exacting process. It requires, first, explicit parental approval and confidentiality of records. Further, all of its categories of eligibility require verification by psychologists, physicians, or other specially licensed personnel. The assessment must necessarily be in the child's native language and take into account the special needs and characteristics of the child. Final judgments cannot be made on the basis of a single test battery administered to all students. Other different assessment techniques must be used to verify or disconfirm initial findings. Final judgments must be made by a team of professionals. Progress evaluations, following the same procedures as the initial diagnostic assessment, must be conducted at least every three years.

Each district is expected to have a program of assessment especially designed for the identification of children who would benefit from special educational services. Further information on the technical aspects of the assessment and the required reporting procedures can be obtained from the Special Education Office of the Nebraska State Department of Education.

Responsibility for the Program

The overall responsibility for the assessment program lies with the district school board and the superintendent. Chief management duties will usually be assigned to various directors of pupil personnel services, assistant superintendents, guidance directors, testing committees, directors of testing, principals and assistant principals, or school psychologists. The assessment examiners are usually counselors or teachers. The personnel having management or administering responsibilities should be encouraged to expand their training and professional involvement in the field of educational measurement.

There is a collective responsibility on the part of all users of the assessment information including the individual student, teachers, counselors, administrators, parents and the public. Each audience of users has a variety of specific needs or uses for the information. Where the audience has difficulty understanding, interpreting, or using the assessment information, they have the responsibility to ask for the changes that will improve its use.

THE GENERAL PURPOSES OF AN ASSESSMENT PROGRAM

There are two general categories of assessment information. In one form, it involves forecasting future performances, predicting, or detecting learning weaknesses. Assessment information so used is diagnostic in value. It is a basis for projecting probable outcomes onto possible future circumstances. The other form assessment information takes is the confirmation of past learning and accomplishment, the measurement of past and present progress. This information, in fact, represents the criterion measures of the school's learning outcomes. It is the variable upon which all instructional and curricular decisions turn since it concerns achievement of the curriculum's goals and objectives. It is also the information by which the school is held publically accountable.

As assessment program collects information on students to record what the students know or can do at a given point in time. The information can be used to mark the passage and progress of students through the learning opportunities. This information, in effect, becomes the criterion measures of achievement. These are the measures of learning by which the teachers, the school, and the student are held accountable. The information can also be used to record what the students are capable of knowing and doing. That is, it is used to diagnose the students' present capability with an eye to how this might affect future learning.

Assessment information may also be used in evaluation studies, and can even be specially collected for such. These studies take the students' performances on criterion measures as effects, and seek to identify what factors within the school's control contribute to the

effect (either in a positive sense or a negative one), or accompany the effect (in a beneficial or undesirable way). Evaluation is a special study in itself and is considered in the Blueprint only with respect to some special points. Both the accounting of student learning by the criterion measure of achievement, and the collection of diagnostic information for the management of learning, precede evaluation and indeed set the occasion for such study. These aspects of assessment will be the central focus of the Blueprint's discussions.

Progress and Passage: The Criterion Measures

The curriculum is divided into units. In each unit, the student encounters a series of deliberately arranged learning opportunities (instruction) and is expected to learn something as a result. If, and how much, the student learns is the central object of measurement. From this information, it is determined if the student is passing, or progressing adequately through the curriculum. This kind of assessment tells what, how much, and when the student has learned.

The criterion measures of passage and progress are either managed by the teacher or by the school. In the first case, the teacher sets the standards of performance that mark the passing grades for the special instruction for which he is responsible. There are conventional standardized tests available for this purpose, but usually this assessment takes place through teacher-made tests. In the second case, the school as a whole furnishes the instruments of assessment, and administers them to the aggregate of the student body, ordered by schools and grades. Although there are many commercially available instruments, the school,

like the teacher, has the option of developing its own instruments.

The school managed assessment of progress includes general achievement or overall educational development as well as special achievement. The passage of curricular units marks special achievements, ones restricted in time and in narrowness of subject matter. The cumulative effect of special achievement, however, can be measured by measures of global achievement or overall educational development. The steady movement on these measures marks general academic progress. The cumulative progress can also be charted by sequences of criterion-referenced measures.

Accountability

The schools manage educational programs. This means they have learners, learning goals and objectives for those learners, and teachers responsible for the instructional strategy and the delivery of the instruction. Different kinds of learning are the outcomes of the programming. An assessment program will consist of the systematic measuring and recording of the learning that results from the programming.

Measures of specific learning outcomes are the first criteria of all educational programming. Higher order judgments of the programming's worth and value are first dependent upon sound method for determining whether or not the specific learning has in fact taken place. The assessment program fulfills the function of accountability if it systematically records the progress of individual students and groups of students on criterion measures of passage and progress, and periodically

summarizes this information. This is simply keeping track of business. A school without accountability procedures is like a business without an accounting department and no annual balance sheets on record.

The cumulative record of the individual student is the basic unit of the procedure. Each student's performance on each criterion measure of passage and progress is recorded in the individual's cumulative folder along with information on the individual's personal and group characteristics, e.g., age and grade. The summarization procedures are no more than group tabulations and averages. For example, the scores of all sixth graders in the district on a reading proficiency test would be averaged and posted in a district summary of this year's sixth grade against the average of sixth graders in past years.

The longitudinal trends obtained in this manner can hold three basic interpretations. First, if the averages slip too many points, deficiencies are noted, prompting special attention to the class as it proceeds farther through the grades, and an examination of possible causes. Second, the average performance can be seen as acceptable and desirable to maintain. Changes in the instructional strategy would be inadvisable. Third, the average performance can be viewed as adequate but leaving room for improvement. Careful changes in the instructional strategy would follow.

The accountability mechanism grounds evaluation efforts in two ways. First, it systematically measures and records criteria which the evaluation design must relate to the possible contributing factors. Second, fluctuations in longitudinal trends set the occasion or need for special evaluative efforts.

Diagnosis

From the students' point of view, two of the principal contributing factors in their learning are their general learning ability and their particular past learnings which serve as special abilities. Their past learning should be detailed and documented by their performances on the criterion measures of passage and performance. The learning abilities that they bring to the learning occasions provided by the educational programming are measured by diagnostic instruments.

Diagnostic assessment is designed to provide information on the students' strengths and weaknesses, their potential achievements, and any learning disabilities. This information serves the school when it identifies students with special instructional needs. It serves the student when it aids in personal, educational, and vocational planning. Four kinds of diagnostic batteries are primarily used: 1) intelligence; 2) aptitude; 3) proficiency; and 4) interest.

Because of the influence of past learning on present and future learning, assessment information on past progress can serve diagnostic purposes for such things as what the student has learned, how much and how quickly. These will set limits on the direction and probable quantity of future learning. There are other factors affecting future learning, however, and those attributable to the individual student (intelligence, aptitude, proficiency, and interest) are usually viewed as the objects of diagnostic assessment. All of these are traits which can be possessed in different measure and degree by each student, and each, when validly measured, provides information on the likely or probable performance of the student in different kinds of future learn-

ing. A sound diagnostic instrument will always specify exactly what kind of future behavior it predicts and show strong empirical correlations with that behavior. As a general rule, diagnostic information should be more conservatively interpreted than the criterion measures because the consequences of error lead to the neglect of talent or miseducation, an unfortunate waste of human resources.

THE VARIETIES OF TESTS AND ASSESSMENT TECHNIQUES

The general purposes of diagnosis and accounting are fulfilled by the combination of those particular tests and assessment techniques administered as part of the assessment program. Each test and technqiue has different characteristics and capabilities. The types of tests and techniques detailed in the following sections will be described by information on the nature of the thing measured, the kind of scores used, and the criteria for evaluating any specimen of the type. Each type will be illustrated by existing tests.

The types are generally classed into three groups: Tests of General Ability; Tests of Special Abilities; and Tests of Typical Performance.[2] The particular categories are listed in Table 1.

TABLE 1
CATEGORIES OF ASSESSMENT INSTRUMENTS

Type	Diagnosis	Progress
General Ability	Intelligence and General Aptitude	Global Achievement or Educational Development
Special Ability	Primary Mental Abilities Differential Aptitude Diagnostic Skill Batteries, & Learning Disabilities	Subject Matter Achievement and Skills Proficiency Batteries
Typical Performance	Interests Attitudes Personality Affective Traits Opinion	

[2]After Cronbach, Essentials of Psychological Testing, Second Edition, pgs. 29-36.

I. GENERAL ABILITY

> "The distinguishing feature of a test of ability is that the subject is encouraged to earn the best score he can."[3]

Measures of general ability assess those abilities which are thought to contribute to learning, thinking and performing in a wide range of circumstances. All are reported in a single, overall score. They are of two principal types: Intelligence (or general aptitude); and Global Achievement (or educational development).

A. INTELLIGENCE OR GENERAL APTITUDE

These measures give indication of what might be reasonably expected of the student. They may be used to classify the student for special instructional attention or for comparison with the student's actual achievement; e.g., is the student doing better than expected? Although many of these measures are frequently called intelligence tests, the accumulation of past learning soon becomes the better predictor of expected achievement. Tests whose content capitalizes on this past learning are more appropriately called "scholastic aptitude" tests.

The scores of these measures are reported on norm-referenced scales. The individual's performance is quantified by comparing it to the performance of a norm group to which the scale is referenced. For instance, an individual's mental age score represents the chronological age group that has a mean performance score similar to the individual's obtained

---

[3]Ibid, pg. 29.

score. The recent trend, however, has been to use percentiles, which follow more statistically lawful methods in reporting overall performance. Note that the intelligence quotient statistics, I.Q., has been used to summarize scores in both approaches. In the classical sense, I.Q. is the ratio of one's mental age to one's chronological age. In the statistical sense, I.Q. is the calculable relationship between an individual's score and those scores of a norm group.

Illustrations: Lorge-Thorndike Intelligence, Otis-Lennon Mental Ability Tests, California Test of Mental Maturity

The test manuals accompanying the tests should provide enough information on the tests' norming procedures for users to judge:

(i) Whether or not it is appropriate to compare their students with those persons in the norm group;

(ii) Whether or not the sample is representative of the norm group;

(iii) Whether or not the sample is large enough.

Evaluation Criteria

The construction of an assessment instrument is complicated. The accomplished result, however, no matter how sophisticated, must meet a few simple standards. If the instrument does, it is probably a good instrument. If correctly applied, it will do the work it is supposed to do. If it does not, even if "correctly" applied, it yields misinforma-

tion, data that is used as if it were true information, but which is actually measuring extraneous and irrelevant things instead of its intended object. But whether a test be good or bad, one must know the reasons why. The standards, met or failed, are the criteria for evaluating the worth of an instrument.

The criteria for judging the worth of any test fall into two camps. The first concerns the quality of the instrument in itself and is discussed in terms of reliability and content validity. The second concerns the work the instrument does and is discussed in terms of various sorts of criterion-related validity -- predictive, construct, and concurrent. Except for reliability, these criteria do not apply equally to all tests and assessment techniques. They will be presented as they become appropriate, depending on the variety of the test or assessment technique. Reliability, because it holds for all, merits general explanation at this point.

One can have a "measuring instrument" that does not measure what it is supposed to measure because it is inaccurate, such as a scale that weighs twenty pounds on the heavy side. When we take the reading from such an instrument, we get misinformation. We think that we weigh 150 when we really weigh 130. Unless one catches onto the constant error, the scale will prove a very unreliable means of telling weight.

When a measuring instrument of educational measurement is unreliable, it means that the score that is observed is probably not a score very close to the truth; it will be off the mark by a large degree of error. The coefficient of reliability is an estimate of just how much error is actually contained in the raw scores which the instrument

yields. The range of values the coefficient can take on run from .00 to 1.00. Instruments with coefficients very close to 1.00 measure what they are supposed to measure almost perfectly. The readings taken from such an instrument would contain a very small margin of error. (See Volume III - Reliability, for the methods for estimating the reliability of assessment instrument.) If the test scores are used to make decisions about individuals, as is usually the case with tests of general intelligence or aptitude, the reliability estimate should be .85 or better.

The effect of error on the interpretation of test scores can be visualized by calculating the standard error of measurement for a test and using this statistic to set confidence bands around the point scores on the scale upon which the test reports its results. The greater the proportion of the scale encompassed by the confidence bands to the total length of the scale, the poorer the measure (see Volume III - Standard Error of Measurement).

The test manual that accompanies the specimen set for a test should always provide reliability estimates for that test and describe the method by which the reliability was estimated. When other things are equal between two alternative instruments, the one with the highest reliability is the better instrument.

The Validity of General Ability Measures: Intelligence tests should serve only diagnostic purposes. The type of validity, therefore, that is relevant to them is predictive validity. Predictive validity insures that the test in fact correlates positively with a future performance of

some note. When the test is used for the classification or selection of individuals for special instructional strategy, this validity should be high. Errors in classification and selection lead to miseducation and must be minimized.

General ability scores also ground judgments concerning the individual's potential, their expected level of achievement given the intelligence scores, and their actual performance. Error here amounts to being misinformed and leads to unrealistic expectations or the neglect of talent.

The predictive validity coefficient should be reported in the test manual accompanying the specimen test, along with details of the empirical methods used to obtain the value. Coefficients calculated from poorly designed and controlled research are spurious. The actual reported value of the coefficient should not be less than .25, though few will exceed .65. Other things being equal, tests with high validity coefficients should be chosen over those having lower validity coefficients.

B. GENERAL ACHIEVEMENT OR EDUCATIONAL DEVELOPMENT

These are general ability tests used to measure progress. They are, in a sense, the future performance which general diagnostic tests are expected to predict. Like other measures of general ability, they are norm-referenced. They tend to draw heavily on content representative of a curriculum, whereas those tests that are diagnostic in purpose will tend to have content representative of abstract psychological abilities. General achievement and diagnostic tests will overlap in content, though,

simply because certain kinds of facility with curricular subject matter tend to be indicative of a certain level of psychological functioning.

Illustrations: Scholastic Aptitude Test, General Educational Development Examination, College Qualification Tests.

Evaluation Criteria

The first question here is: Is there really progress? High reliability estimates are necessary to confidently answer this question.

The second question is: What has been accomplished? Answering this involves construct and content validity. If the school is willing to assume that there is a valid psychological construct called "general educational development," then it should expect the test under question to show positive correlations with other tests that also purport to measure such a construct. This is concurrent validation and studies establishing this may be reported in the test manual.

The actual items of the test in question can be examined from two points of view. First, if the school accepts the notion of a general construct of educational development, the items of the test will be evaluated for their representativeness of this general construct. Second, if the school wishes to measure as accurately as possible the students' mastery of the curriculum that it is teaching, then the items are evaluated for their representativeness of the actual curriculum of the school. In the second case, the content validity of the test must

be established by procedures appropriate to the local district.

A last question, seldom asked, is: Of what value is the accomplishment? Empirical studies showing that something good or useful happens to students because they have mastered this or that part of a generalized curriculum are not likely to be reported in the test manual, but could be conducted locally to significant effect.

II. SPECIAL ABILITIES

Most of the measurement effort in assessment is devoted to special abilities, either for diagnostic purposes or to mark progress. Indeed, almost all measures of general ability break down the items of the test by special abilities and report a profile of part scores. The field of special abilities also presents the greatest variety of measurement technology, including conventional standardized tests, teacher-made tests, and several kinds of criterion-referenced measurement.

A. PRIMARY MENTAL ABILITIES AND APTITUDE BATTERIES

The first place where special mental abilities tend to be indexed is on general intelligence or aptitude tests which report part scores for verbal and numerical abilities. The evaluation of these tests is identical to that of the parent tests except that special attention needs to be given to the reliability and validity of the part scores, should they be used extensively in the interpretation.

The advent of the statistical techniques of factor analysis has led to the refinement of the conceptions of primary and special mental

abilities. Notable efforts include: (1) Thurstone's six primary factors (verbal, number, spatial, word fluency, memory, and reasoning[4]); and (2) Guilford's structure of the intellect (memory, cognition, convergent thinking, divergent thinking, and evaluation[5]). The principal uses of these tests remain in research settings. Other theoretical perspectives on this search for particular mental abilities include Piaget's theory of formal operations, Bloom's taxonomy of the cognitive domain and Gagne's learning hierarchies, though practical instrumentation for measuring these is not widespread.

The Differential Aptitude Tests and the General Aptitude Test Battery, on the other hand, have been most thoroughly validated in applied settings. These tests do not attempt to measure pure abilities, but rather measure abilities that are complexly related to future performance in different occupations or educational programs. The separate scales are designed to be as independent of each other as possible. The scales upon which the point scores are reported are usually norm-referenced. It is especially critical to have an accurate idea of the norm group when making interpretations of the point scores.

The principal use of differential aptitude scores is in guidance and and counseling, where the individual is presented with the results as a fair picture of his abilities on a particular set of measures. It is very important that the students understand that the scores have only

---

[4]Thurstone, L. L. Primary Mental Abilities. Psychometr. Monogr., 1938, No. 1.

[5]Guilford, J. P. Fundamental Statistics in Psychology and Education (3rd Ed.). New York: McGraw-Hill, 1956.

probabilistic relationships to future performance and that many other factors, measured or not measured, will ultimately contribute to their success or failure once they choose a set of goals to pursue. Factors such as interest, motivation, discipline and attitude may overcome the limitations suggested by only moderate aptitude, or the same factors may operate to cancel out strong aptitudes. It is especially helpful in counseling if the scales are explained in terms more easily understood by the students.

Illustrations: Primary Mental Abilities Tests,
General Aptitude Test Battery
Differential Aptitude Tests.

Evaluation Criteria

The testing manual accompanying the test specimen should report the reliabilities of each scale of the battery, their intercorrelations, and their individual predictive validity coefficients.

Overall reliability estimates and validity coefficients will usually be higher than the part score estimates and should not be used to calculate the standard error of measurement for the part scores. The intercorrelations between part scores should be low. Where they are high, the two correlated parts are apparently measuring the same thing, and consequently provide redundant information which can be misleading in forming interpretations. One or the other can be discarded. The need for good predictive validity for each part score should be obvious.

Undesirable consequences of miseducation, unrealistic expectations, or neglected talent arise from being misinformed about special abilities as well as from being misinformed about general abilities.

Reading Readiness: The reading readiness test is like the preceding special ability measures in that it is principally used to assess the prospects of future performance at the onset of reading instruction. These are obviously not measures of reading skill or progress in reading skill and so essentially serve as a screening device. The evaluation criteria include again reliability and predictive validity. Reading readiness scores should have a reasonably high correlation with later reading achievement or they are not performing the work they are intended to perform.

B. SPECIAL ACHIEVEMENT

There are four major approaches to the measurement of special achievements. These are represented by: (1) conventional standardized tests covering particular subject matter areas or basic skills; (2) common types of criterion referenced tests (objectives-referenced, content-specified, or domain-referenced); (3) teacher-made tests for classroom achievement; and (4) the techniques of applied performance testing.

(1) Conventional Standardized Tests

When developed along the traditional divisions of subject matter, these tests frequently occur as a battery or a series. The content

areas are such familiar things as history, general science, literature, etc. Other batteries will measure more directly the growth in individual skills. In all cases, these sorts of tests essentially represent the criterion measures of particular learning outcomes as opposed to the more global indices of educational growth.

The scales are typically norm-referenced, although cut-off scores may be set upon the raw score numberline as specific performance standards. The reported point scores are usually percentiles. Both the subject matter area and the skill area tests may serve as the criterion measures when systematically reported as part of the accountability mechanism. They also serve well as the criterion measures in evaluation designs. The skill area tests, however, frequently cover enough specific and independent skills to serve diagnostic purposes. Inasmuch as they provide accurate measures of a student's proficiency with the specific skills, they assess capabilities which will affect future performance. Consequently, they can be useful in identifying special instructional needs or in counseling students in planning future education or a career.

Illustrations: Stanford Achievement Test, Iowa Test of Basic Skills.

Evaluation Criteria

The questions of the tests' reliability and the adequacy of the norm groups are important with these tests, as with any conventional standardized test. When used for diagnostic purposes, the predictive

validity of the part scores is important. When used to measure special achievement as a criterion, the content validity of the tests is central.

Establishing the content validity of the test is a logical process that must be carried out with a knowledge of the local curriculum. Although, in a general sense, the content of these tests never strays far from a plausible representation of the content area, they are designed to be generalized, representing many possible curriculums, and to foster a broad distribution of obtained scores. Consequently, the same items may not be representative of a particular curriculum and on the whole, fail to reach the proper alignment with the curricular subdivisions that are locally emphasized. Further, the manner in which the items are cast may also reflect a style of problem statement that is different from the manner in which the students' teachers actually pose questions. Too much mismatch between the test's content and style and the school's curriculum and teaching style will yield vastly misleading results.

The full process for establishing a test's content validity for a local district should ideally proceed as follows.

(i) Descriptive Review of the Curriculum:

A particular curriculum will be constructed around certain concepts or categories of instruction. Within each category, the more specific objectives delineate what is taught. The objectives themselves can usually be broken down further. When this analysis is pursued and written out in the terminology of the local district, the result will be an inventory of the curriculum. It is the content of this inventory, in the areas to be assessed, that must be

matched against the test's content to establish content validity.

(ii) Matching:

A representative committee of school personnel should examine each item of a test against the inventory, marking where each falls, and judging the appropriateness of the item's style, given knowledge of the student body and the methods and manner of instruction. The distribution of the items over the categories of the inventory should be displayed graphically. There are no accepted quantitative standards for marking the threshold of inadequate matching. The committee, however, can reach consensus judgments which reflect how much confidence can be placed in judgments as to how much of the curriculum students are really mastering. Does a high score probably mean a high degree of mastery of the local curriculum? Could a student master large portions of the local curriculum yet still obtain a low score on the test because of imbalances between contents? Unless a high level of confidence can be reached, the test as a criterion measure of the learning outcomes is unfair, both to the students and to the teachers.

It is possible that the tests' content is judged to be more representative of a subject area than the current curriculum is judged to be. In this case, the curriculum would be wisely redeveloped and the test, in effect, becomes one of the specifications of the curriculum design. Inasmuch as the tests' content does seem to validly represent a general construct in subject matter, this endeavor is creditable. As a rule of thumb, however, all objects of assessment should be grounded upon conceptions of learn-

ing outcomes that are articulated by the curriculum. The principal exception to the rule would be special purpose assessment intended to investigate special areas of concern or the possibility of special instructional needs.

(2) Criterion Referenced Measurement

The principal difference between criterion-referenced tests and norm-referenced tests resides in the manner in which the scale of reported results is constructed. A norm-referenced score compares an individual's performance relative to the performance of the norm group. This sort of comparison is often very useful, but lacks any direct interpretation of how much of the content area that the test assesses was actually acquired. For instance, how much content did the norm group really acquire? A criterion-referenced instrument, on the other hand, attempts to formulate the individual's score in a direct relationship to the actual amount of content mastered. This is usually accomplished by designing the test's content to be strictly representative of the larger content area to be assessed. If this content validity is assured, then a cutting score is set as the threshold at which most reasonable persons would be confident that most, or a sufficient amount, of content mastery is indicated. If the cutting score is surpassed, the individual is judged to have acquired, or mastered, the content area under question. When the content area represents a domain of skill applications, the test becomes a skill proficiency test. The threshold score then marks the point at which one can reasonably view the student's proficiency as reliable and strong.

Defining the content area in constructing a subject matter measure may take several forms. The most general form would be content-specified. The actual content area in this case would be too indefinitely large to be exhaustively listed. It can, however, be logically broken down and represented by a table of specifications. The items of the tests would be expected to fit the specifications so outlined.

When the content is specified by objectives, the test is objective-referenced. Since objectives tend to rather tightly limit the content, tests of these sorts are relatively short, usually five or seven items to an objective. One major advantage of these sorts of tests is the flexibility they offer. Each objective is viewed as an independent unit of content. A system of these objectives would be expected to provide comprehensive coverage of a large content area, third grade reading skills, for instance. A district would have the option of selecting only those objectives that match with instructional objectives in the curriculum.

Objectives-referenced tests readily lend themselves to the formative control of the individual's progression through the sequence of learning opportunities provided by the school. The student must demonstrate mastery of an objective before venturing on to the learning opportunities provided for the next objective in the sequence. Overall progress can be gauged as a function of the number of objectives mastered in a given period of time.

When the universe of content for a particular test can be exhaustively detailed (e.g., all one-digit subtraction problems), the test is domain-referenced. The items of the test may randomly sample the domain

(50 out of 100), or represent the domain completely (100 out of 100). Because of the more stringent requirements of definition, domain-referenced tests have the same narrowness that objectives-referenced tests have, although the length of the test will frequently exceed 10 items. Its uses are also in passage and progress assessment, with the possibility of special use as a control on the progression through the learning opportunities.

If a battery of criterion-referenced tests is administered all at once, in any combination of content-specified, objective- or domain-referenced, the composite profile may be used for diagnositc purposes. This data will often figure in the student's initial placement in the sequence of learning opportunities. This is especially true of skill diagnostic inventories in reading or mathematics.

Illustrations: Nebraska Skills Tests for the Assessment of Reading, Nebraska Skills Tests for the Assessment of Mathematics.

Evaluation Criteria

Any measurement device must prove reliable; however, the frequently brief nature of most criterion-referenced tests has altered the approaches estimating their reliability. Internal consistency methods have proven useful. The test manual accompanying a system of criterion-referenced tests should provide information on this question.

The most critical evaluative criteria, however, is content validity. This should be established through the same method recommended above

with respect to conventional standardized tests of special achievement.

(3) Teacher-Made Tests

The most important and the most frequent form of assessment that the student faces is the teacher-made test. It is usually designed to gauge the student's achievement in a course of study within the confines of the classroom unit. These tests most resemble criterion-referenced tests except that the teacher does not have the luxury of intensive development work on the tests and typically uses graded levels of performance for the scale in which the scores are reported. They can, however, be methodically constructed.[6]

A teacher should be just as aware of the methods and techniques of measuring what they teach as they are of curricular developments occurring in the subject area they teach. A district's assessment program can also support the teacher's efforts by operating test scoring services and by assisting teachers in developing and organizing item pools.

(4) Applied Performance Testing

When the thing measured is conceived more as a kind of doing or behaving rather than as a content area or subject matter knowledge, the kind of measurement technology alters and is properly called applied performance testing.[7] The results are understood in terms of what the

---

[6]Several excellent guides are available as a package from the Educational Testing Service, Princeton, New Jersey: Making The Classroom Test: A Guide for Teachers; Multiple Choice Questions: A Close Look; and Short Cut Statistics For Teacher-Made Tests (by Paul B. Diederich).

[7]Performance testing without the modifier "applied" traditionally refers to the measurement of non-verbal behavior. Applied performance measurement has no such restriction.

subject can do or is capable of doing. Consequently, this kind of measurement is more appropriate in the assessment of skills (e.g., essential, functional, athletic), and of specific learning propensities or disabilities. These techniques can be norm-referenced in some cases, where extreme rank is sufficient to detect the presence or absence of a condition or property of the individual, aphasia, for instance. But these cases would all be diagnostic, where criterion related validity is pre-emptive. Where the performance scale is criterion-referenced and used as a benchmark of progress, the construct validity is foremost. Does the behavior required to perform the assessment task clearly signify the presence or absence of an ability possessed by the individual?

Applied performance tests are demonstration oriented, and as such are ideally suited for competency measurement where competence is viewed as the applications of a skill or of a cognitive process. Discussions of the application of applied performance testing in school setting can be found in "Applied Performance Testing in the Classroom," a paper available from James R. Sanders, The Evaluation Center, College of Education, Western Michigan University, Kalamazoo, Michigan 49008, and the Administrative Manual of N-ABELS, available from School Management Services, Nebraska Department of Education.

A. Performance Sampling

The best measure of a behavioral ability is the direct observation of the performance of an individual under task and stimulus conditions which call for a behavior representative of the ability. If one wants to know if students can fix a radio, give them broken but repairable radio sets and watch to see if they can fix it. In principle, one can observe the

individual's behavior in the process of repairing or one can observe the end product, the radio allegedly repaired. In one instance, a behavioral process is under review, where one can assess the efficiency with which the radio is repaired, the appropriate and adept use of tools, and the thoroughness of the procedure in sum. The other instance is product evaluation: the set works or it doesn't; the connections are smoothly soldered; etc. In both cases, process and product, the individual may create the observation record, or trained observers can create the observation record, or both kinds of records can be produced. The scale in which the individual's score is set derives from these observation records, e.g., answer sheets, time charts, checklists, and rating scales.

Evaluation Criteria

The individual-produced record of observation tends to be the more objective and should have excellent test-retest reliability. The subjectivity of the observer-produced record of observation can be controlled by observer training and indices of inter-observer consistency.

The approach to the construct validation differs between diagnostic uses and progress benchmark uses. In both cases, the thing measured must be given a psychological interpretation; that is, what is measured must be viewed as a distinct behavioral capacity or disability. The item or tasks must be indicative and representative instances demonstrating the capability. If the information, however, is used as diagnostic, the measure must be shown to have an empirical relationship with other conditions which significantly shape one's future performances. When the information is used as a mark of progress, the value of the thing accom-

plished should be evaluated. How often will the individual use the capabilities acquired, to what ends, and for what benefit? Thus, even with a progress-criterion, there is a need to demonstrate that progressing to such and such a point will mark a difference in such and such future circumstance.

B. Simulation

Where the stimulus conditions are selected to represent the more essential features of more complex situations that would be impossible to replicate, the sample of behavior is observed under simulation. This is indirect measurement. Both individual-created and observer-created observation records of the sample are possible, but the distance between the sample and actual behavior defined as the capability is lengthened. There are two approaches to evaluating the effect of the simulation. The first is to assess the logical consistency of the behavior performed in the simulated setting with the behavior that would be performed in the actual setting. The more features in common, the more confident one can be of the ability to perform competently in the actual setting. The more features that are missing in the simulation but which are important to the actual setting--genuine motivation, real money, crisis reasoning, etc.--the less confidence one can have that transfer will occur. The second approach involves the same logical analysis, except that the difference between the stimulus conditions of the simulated setting and actual setting are compared, instead of the behavior performed. This is to assess the authenticity of the simulation, how much like the real setting the simulated setting is. Again, the more in common, the

greater the confidence in transfer; the more missing features, the less the confidence. At first, this is a logical question requiring logical method. The question of transfer is open to empirical inquiry and, consequently, open to criterion-related validation.

C. Applied Performance Batteries

Several individual applied performance tests are often united for the purposes of assessing a broad spectrum of behavioral abilities, for example, those representative of the application of basic skills. They may also be brought together into a sequence that represents kinds of learning that progress through increasingly complex levels of accomplishment. Each individual assessment task should be evaluated separately, according to the kind of criterion-referenced measurement involved.

Illustrations: Nebraska Assessment Battery of Essential Learning Skills, The Denver Developmental Screening Test.

III. TYPICAL PERFORMANCE

The typical performances of a student represent those behaviors which the student is quite apt to show in a given setting rather than the best behavior of which the student is capable. For instance, a student may be capable of comprehending Joyce, Goethe, or John Locke, yet typically read science fiction when not required to otherwise. The measurement of typical behavior is quite undeveloped in educational practice, mostly because the objects of measurement, usually affective

traits or ordinary habits, present difficulties in both conceptual and technical precision that are not found in the measurement of ability. Since affective and habitual behavior seem more intimately associated with personal identity than are abilities--the last bastions of identity if indeed one's abilities are utterly undistinguished--the rights of individual privacy have also limited the growth and spread of the measurement of typical performances.

The categories of typical performance measures include habitual behavior, attitude, interest, and personality. The assessment of interests is the most commonly practiced and is usually carried on solely for the benefit of students making future career and educational plans. Of the others, attitude measures are often used in evaluation, personality measures in counseling, and habitual behavior measures seldom, if ever. Only interest will be discussed here.

Interests Inventories

There are three principal approaches to the construction of interest inventories, each lending itself to an interpretation different from the others. The first approach involves comparing the individual's pattern of choices on the inventories with keys that describe the patterns exhibited by specially defined groups, e.g., engineers or forest rangers. If an individual's interest patterns match that of any given group, then he would be expected to fit into that group, however it is defined. The Strong Vocational Interest Blank was developed using this approach.

The Kuder Preference Record was developed by using a second approach, that of statistically analyzing the items and grouping homogeneous items together into a scale. The scale would then represent an interest area like music, outdoor, clerical, etc.

The third approach involves grouping items according to a scheme of logical classification, without the aid of statistical and empirical techniques. An example of this approach is the Occupational Interest Inventory by Lee and Thorpe.

The Strong and Kuder instruments have been especially well studied. The choice between them rests mainly on one's preference for the approach that was used in their construction.